UiT

THE ARCTIC UNIVERSITY OF NORWAY

TROLLing

defining, building, and operating an open archive for linguistic data

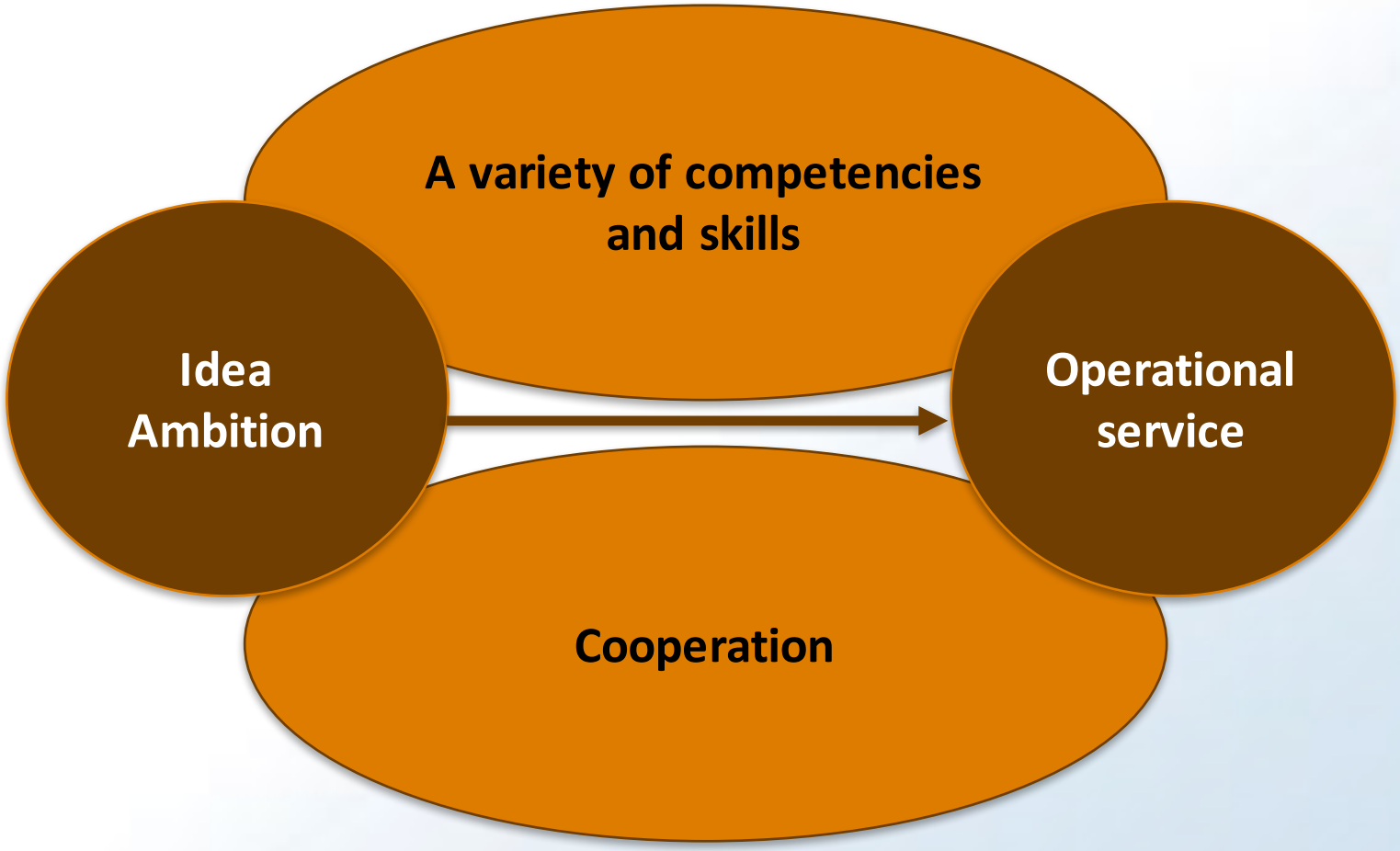
Helene N. Andreassen, PhD

UiT The Arctic University of Norway

2nd Workshop on Standards for Data Citation and Attribution in Linguistics

University of Texas at Austin, April 8-10, 2016

Outline


A variety of competencies and skills
Idea Ambition
Operational service
Cooperation

UiT and open access

*UiT will be recognized by a culture for active dissemination through **open channels for publishing**, as well as through exhibitions, journals and the media.*

UiT strategic plan 2020

Digital – above all!

Photo: Rune Ytreberg

Main ambition Explore and develop the digital possibilities, and use these to strengthen our services to employees and students

Strategy Take a central position in the work with archiving and dissemination of research data, locally and nationally

University Library strategic plan 2020

How TROLLing came to be

1. Inquiry in 2013 from the UiT linguistics community turns the library's ambition to work on open research data into action

2. Working group put together, consisting of researchers and subject specialists in linguistics, OA specialists and system developers

3. Establishment of a three-member scientific advisory board

4. Development guided by scientific needs and international solutions

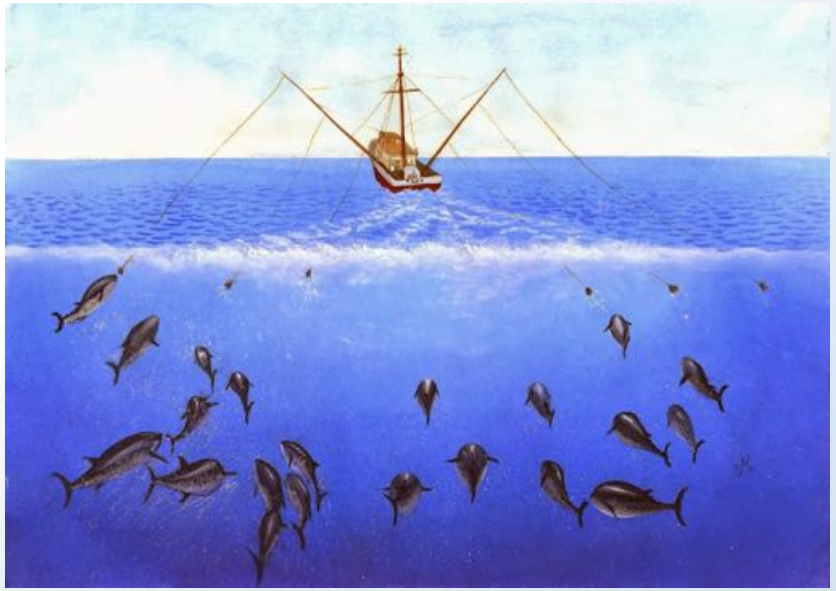
5. Launch in June 2014

Photos: flickr.com/photos/kimgskytte

TROLLing
The Tromsø Repository of Language and Linguistics

Archive for open linguistic data and statistical code

- International service, open to researchers across the world for upload and download
- Maintained and curated by the University Library
  - Relevance of uploaded data
  - Quality and comprehensiveness of metadata
  - Description and format of uploaded data



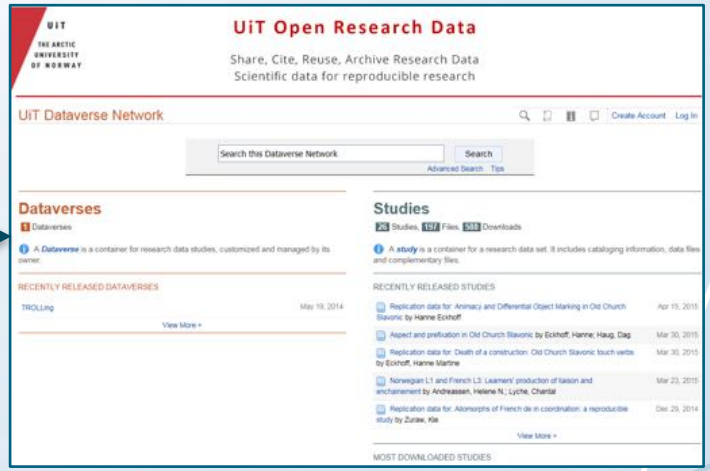
opendata.uit.no

The platform

Guiding principle: Be future-oriented, and think bigger!

TROLLing built on the Dataverse platform (https://dataverse.harvard.edu/)

- Allows adding of other types of datasets using the same tools and templates
- Facilitates harvesting of data by international services
- Complies with DataCite (https://www.datacite.org/)



Adaption of metadata template

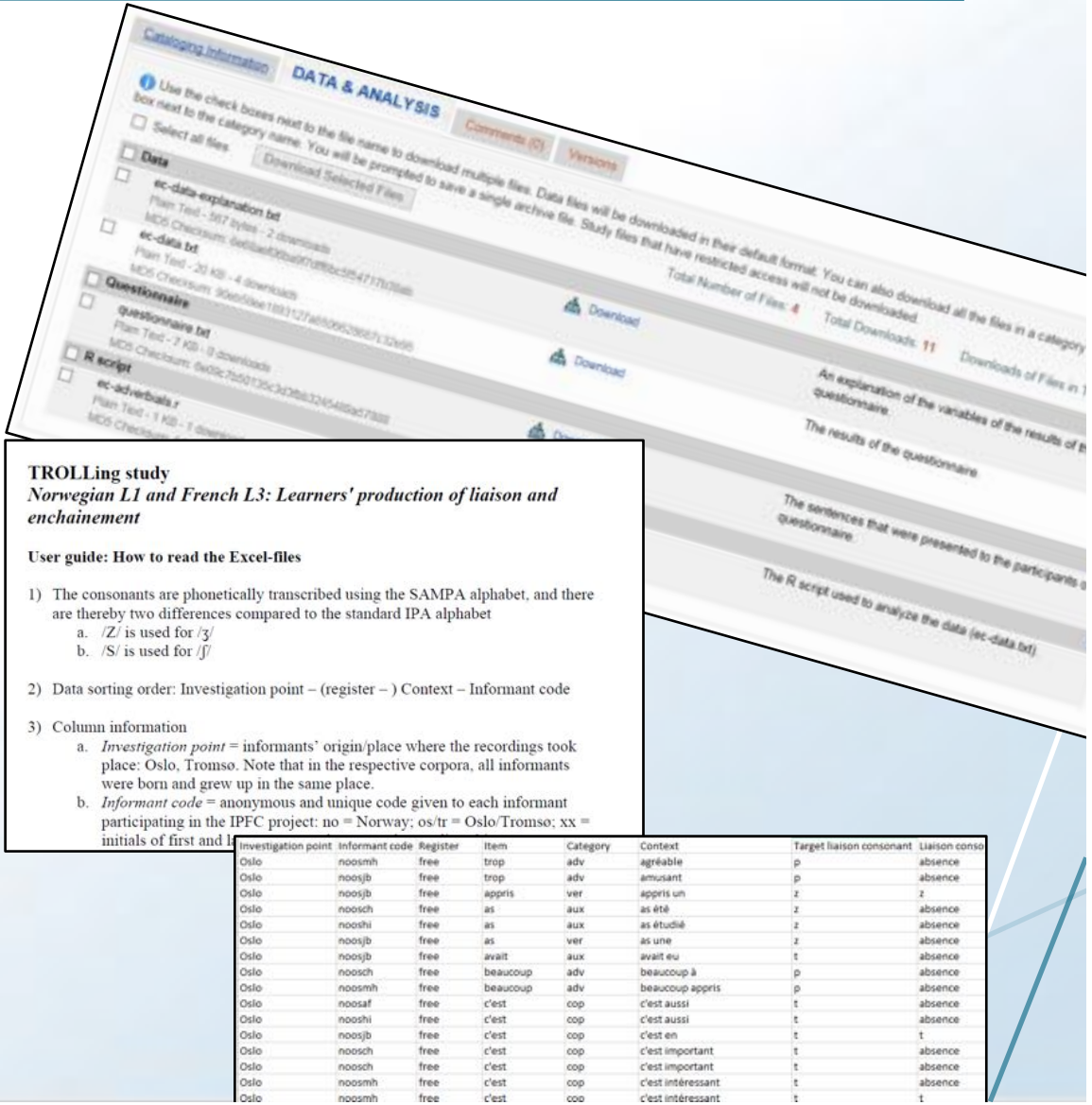
How to optimize retrieval of data

- Topic specification
  - Field
  - Time-depth
  - Topic
  - Free-text keywords
- Description
  - Abstract
  - File content

Setting of requirements
How to optimize reuse of data

- Description
  - In template
  - In read-me file
  - In data file (column headings, file name, etc.)

- Persistent file format
  - Non-proprietary
  - Open
  - Standard character encoding (UTF-8)



Update to come

Dataverse Version 4

- Recently released
- Trolling in the process of being migrated

- Important improvements
  - Richer and more flexible metadata template
  - Tagging on file level, improving the search function
  - Improved metrics: views, downloads, citations, shares

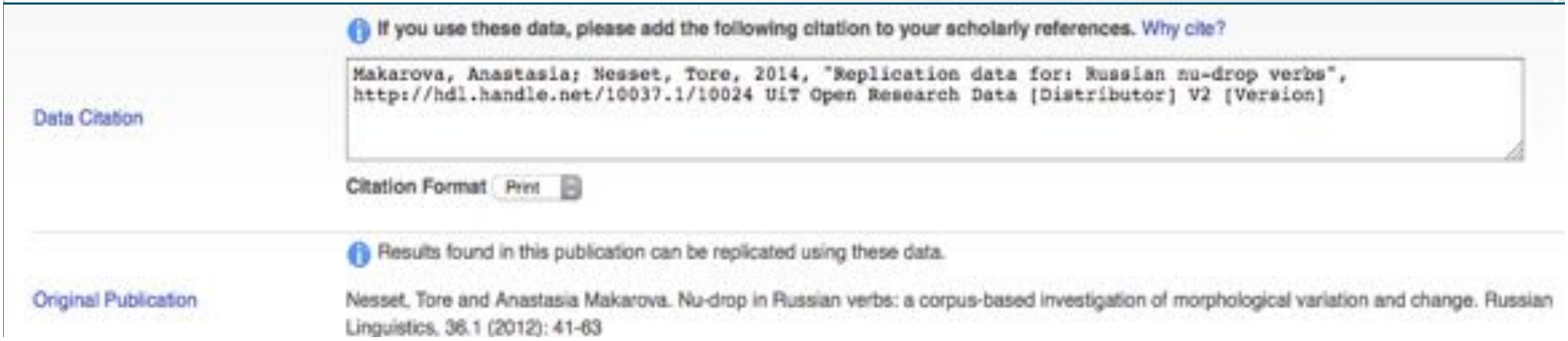
Citing the data

Built-up of dataset citation

- Persistent identifier
  - Doi shortly available
- Data description
  - “Replication data” or other
- Version indicator
  - Previous versions accessible

Requirements on reuse of data

- Standard license selected: CC0
  - Meet the potential problem of attribution stacking
- Citation in line with good academic practice
  - Use the reference as provided
  - (Add subset info if appropriate)

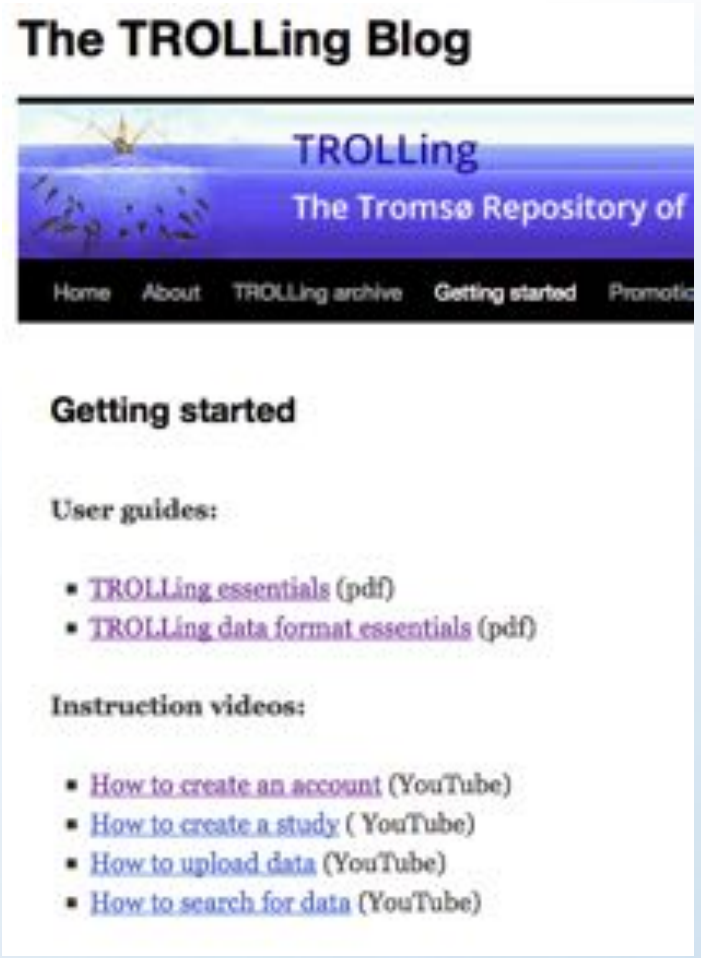


TROLLing
The Tromsø Repository of Language and Linguistics

Archive for open linguistic data and statistical code

- International service, open to researchers across the world for upload and download
- Maintained and curated by the University Library at UiT
- Assisting and educating the users
  - User guides
  - Instruction videos

  ➢ Blog interface for communication
  ➢ Cooperation with faculty



site.uit.no/trolling

TROLLing
The Tromsø Repository of Language and Linguistics

Archive for open linguistic data and statistical code

- International service, open to researchers across the world for upload and download
- Maintained and curated by the University Library at UiT
- Development of user guides and promotional material in cooperation with faculty
- Marketing in every channel possible
  - Promotion material
  - YouTube

  ➢ Cooperation with faculty, graphic designers and video producers

https://www.youtube.com/watch?v=uEf0c0NT9_A

Outreach

Conferences and meetings: presentations and workshops

Laura Janda

- *Slavic Cognitive Linguistics Conference*, U. of Sheffield and Oxford, 2015.
- *13th International Cognitive Linguistics Conference*, Northumbria, 2015.
- *Palatalisation Workshop*, CASTL/UiT, 2014.

Helene N. Andreassen

- *Journées FLOraL-PFC: PFC dans le champ phonologique*, Paris, 2015.
- *Journées FLOraL (Français Langue ORAle et Linguistique)*, Paris, 2014.

Philipp Conzett & Leif Longva

- *emtacl15 - emerging technologies in academic libraries*, Trondheim, 2015.

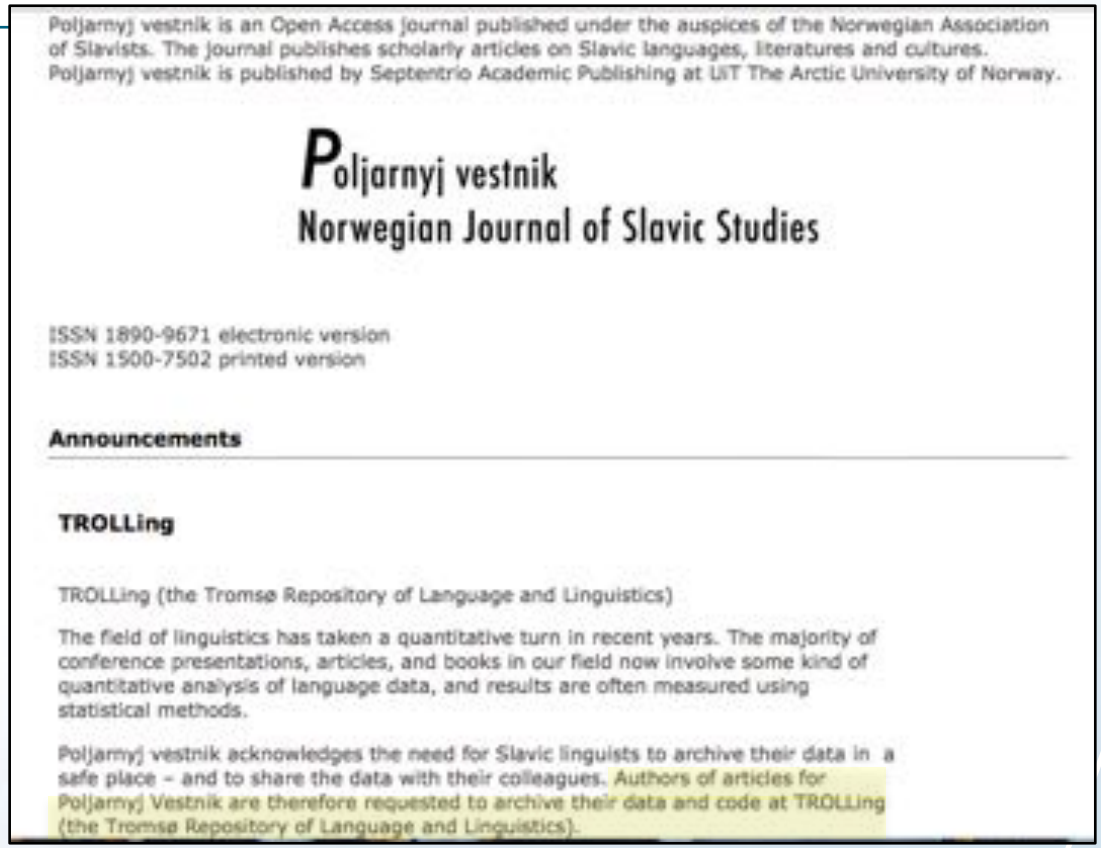
Philipp Conzett & Obiajulu Odu

- *Dataverse Community Meeting*, Harvard, Cambridge, MA, 2015.

Outreach
Approaching the publishers

- Encouraging from above
  - Journal editorial boards
    - put TROLLing into guidelines
  - Cristin (Norwegian National Research Information System)
    - create a category “data collection” or “dataset”/make it count

- Encouraging from below
  - Networks (TROLLing team and UiT linguists)
  - UiT based journals
  - OJS-Dataverse plugin (TBT)
  - Individual projects



Poljarnyj vestnik is an Open Access journal published under the auspices of the Norwegian Association of Slavists. The journal publishes scholarly articles on Slavic languages, literatures and cultures. Poljarnyj vestnik is published by Septentrio Academic Publishing at UiT The Arctic University of Norway.

Poljarnyj vestnik
Norwegian Journal of Slavic Studies

ISSN 1890-9671 electronic version
ISSN 1500-7502 printed version

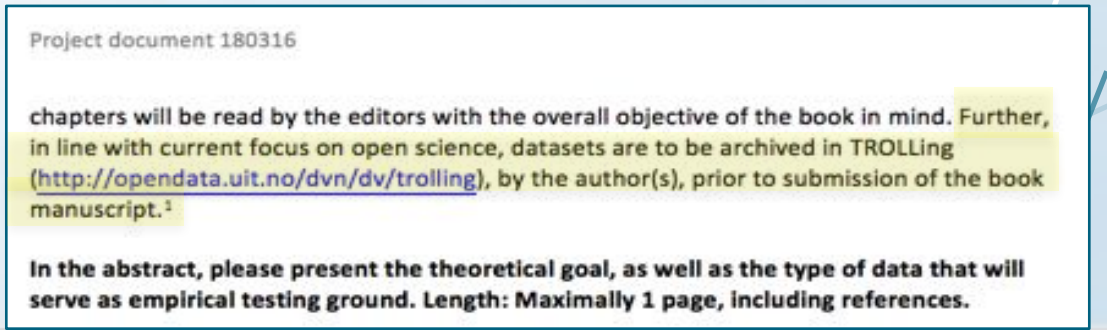
Announcements

TROLLing

TROLLing (the Tromsø Repository of Language and Linguistics)

The field of linguistics has taken a quantitative turn in recent years. The majority of conference presentations, articles, and books in our field now involve some kind of quantitative analysis of language data, and results are often measured using statistical methods.

Poljarnyj vestnik acknowledges the need for Slavic linguists to archive their data in a safe place – and to share the data with their colleagues. Authors of articles for Poljarnyj Vestnik are therefore requested to archive their data and code at TROLLing (the Tromsø Repository of Language and Linguistics).



Project document 180316

chapters will be read by the editors with the overall objective of the book in mind. Further, in line with current focus on open science, datasets are to be archived in TROLLing (http://opendata.uit.no/dvn/dv/trolling), by the author(s), prior to submission of the book manuscript.[1]

In the abstract, please present the theoretical goal, as well as the type of data that will serve as empirical testing ground. Length: Maximally 1 page, including references.

Outreach
Visibility in social media

Facebook

– where “everything now happens”

- Updates
  - New uploads
  - Presentations/workshops
  - Technical information
- Collaborative management
  - TROLLing curators
  - Faculty research assistant

User activity in TROLLing (per April 7, 2016)

Numbers

- 40 studies
- 1394 downloads
- 105 registered users
  - 19 countries
  - Europe, Asia, North- and South-America

Associated publications

- OA journals
- Paid journals
- No publication

- PhD thesis
- Master thesis

Contributors

- 24 unique contributors
  - 5 countries
  - Europe, North-America

User activity in TROLLing (per April 7, 2016)

Content

- Subfields
  - Semantics, syntax, morphology, phonology, phonetics
  - Synchronic, diachronic, first and second language acquisition
- Languages
  - Czech, Old Church Slavonic, Russian, Ukrainian
  - French, Romanian, Spanish
  - German, Norwegian
  - Saami

Content

- Types of data
  - Tables, charts
  - Audio, video
  - Scripts, experimental method

Fully operational service and why curation is still necessary

TROLLing identity, a clear definition It is a place for open, structured datasets belonging to the science of langue

- Yes
  - Structured, well described, openly accessible datasets
- No
  - Metadata only
  - Primary data
  - Sensitive data
  - Bibliographies, dictionaries, national anthems
  - (To be continued)

«Guidelines so easy that my grandmother would not have any problems uploading data.»

Member
TROLLing Scientific Advisory Board

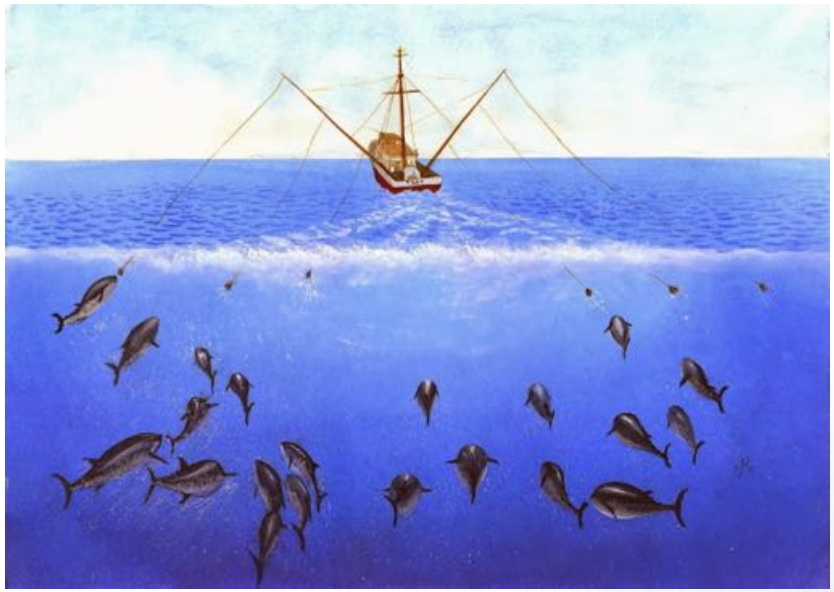
BUT

- Researchers have little time
- Researchers are not used to think about data management

VIA CURATION

- ➢ Assistance and training
- ➢ Consistent optimization

To learn more about TROLLing



- Visit the archive at opendata.uit.no
- Visit the blog at site.uit.no/trolling
- Contact us at trolling@ub.uit.no
- **New idea** Join us in a TROLLing webinar, where we can have a look at the archive together, live, all while being located in different parts of the world

UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Thank you for your attention*

Helene
helene.n.andreassen@uit.no

TROLLing
trolling@ub.uit.no

*Thanks to Philipp Conzett, Stein Høydalsvik, Laura Janda, and Leif Longva (UiT) for useful information and constructive comments

Tromsø, April 3, 2016 (private photo)